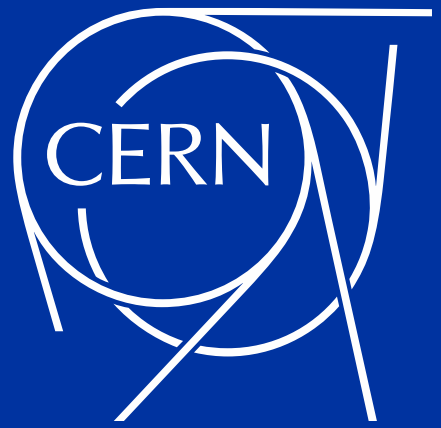


EOS 5: New features & future plans

Abhishek Lekshmanan on behalf of the EOS team

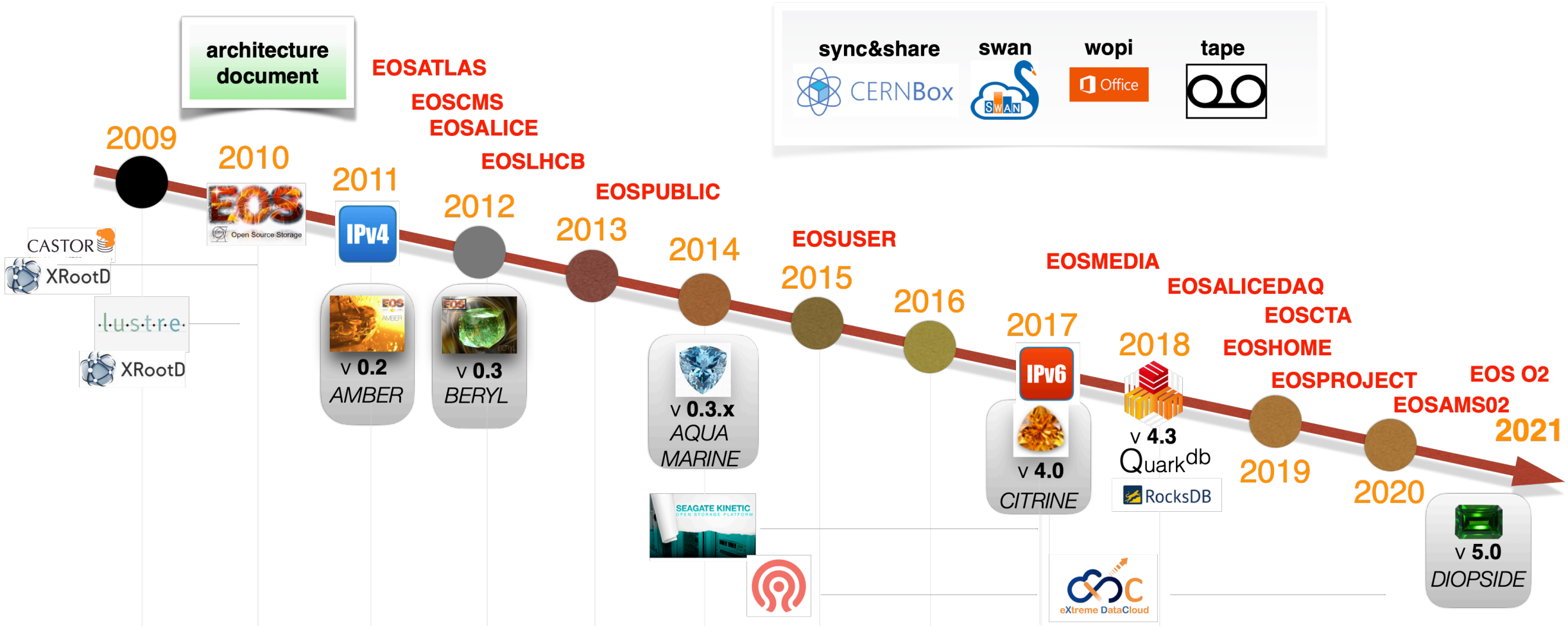
HEPIX Autum 21

Table of Contents

Introduction - What is EOS?

- Open Source distributed storage system designed & developed by CERN IT
- Filesystem like & HTTP interfaces - FUSE, XROOTD, HTTP, CIFS, WebDAV
- HA, fault tolerant & reliable file storage - Erasure Coding and/or Replication
- Multiple authentication methods - KRB5, X509, shared secret, tokens, OAuth2
- Built on top of XROOTD framework

Timeline


architecture document
EOSATLAS
EOSCMS
EOSALICE
EOSLHCB
EOSPUBLIC
EOSUSER
EOSMEDIA
EOSALICEDAQ
EOSCTA
EOSHOME
EOSPROJECT
EOSAMS02
EOS O2
sync&share
CERNBox
swan
SWAN
wopi
Office
tape
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
CASTOR
XRootD
l·u·s·t·r·e
XRootD
Open Source Storage
IPv4
v 0.2
AMBER
v 0.3
BERYL
v 0.3.x
AQUA MARINE
SEAGATE KINETIC
IPv6
v 4.0
CITRINE
v 4.3
Quarkdb
RocksDB
eXtreme DataCloud
v 5.0
DIOPSIDE

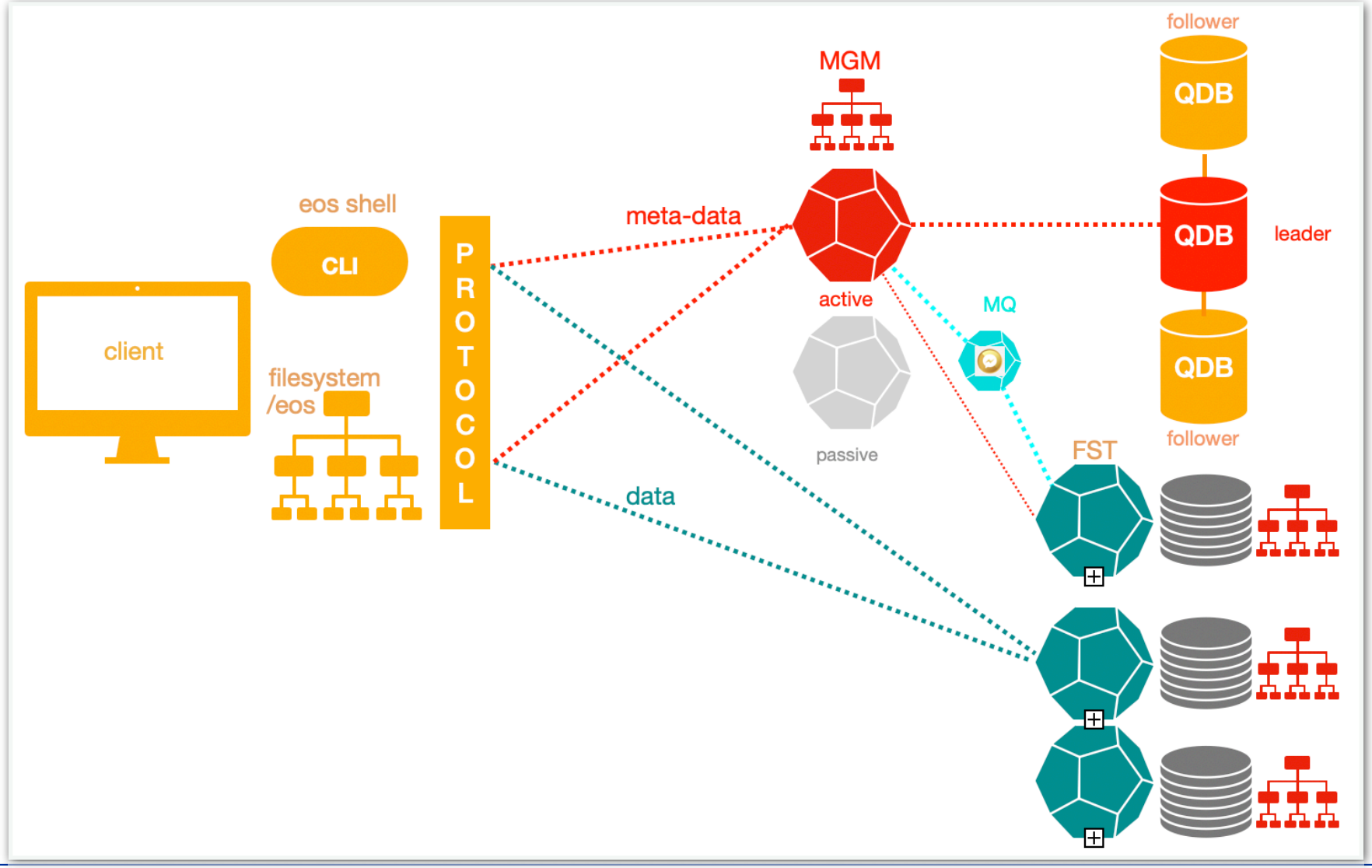
Architecture

- MGM: Metadata Server, runs the main namespace, actual data persisted in QuarkDB, LRU caching on prefetching metadata, does authentication & authorization
- QuarkDB: HA Replicated persistent KV store
- MQ: Message Queue between FSTs/MGM
- FST: File Storage Servers, run on JBODs, replication/EC layouts possible

Architecture

Components

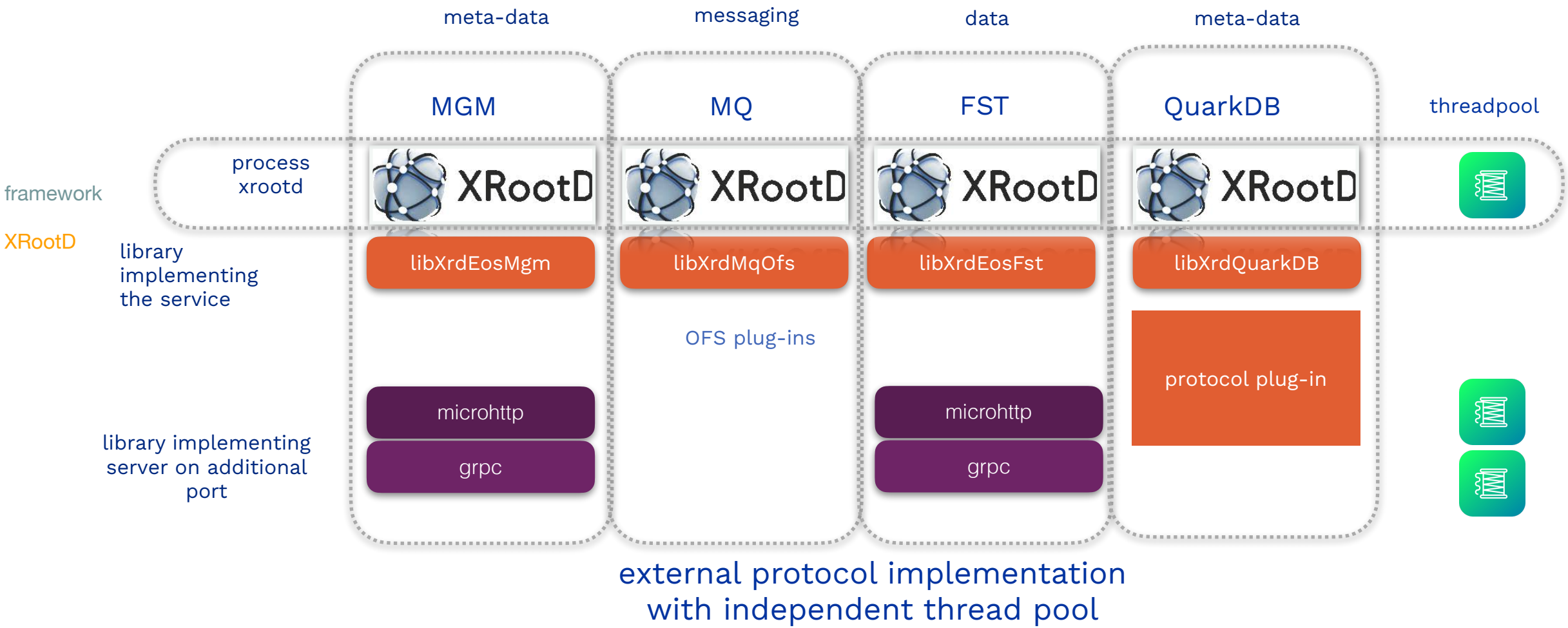
MGM
MQ
FST
QDB
Clients



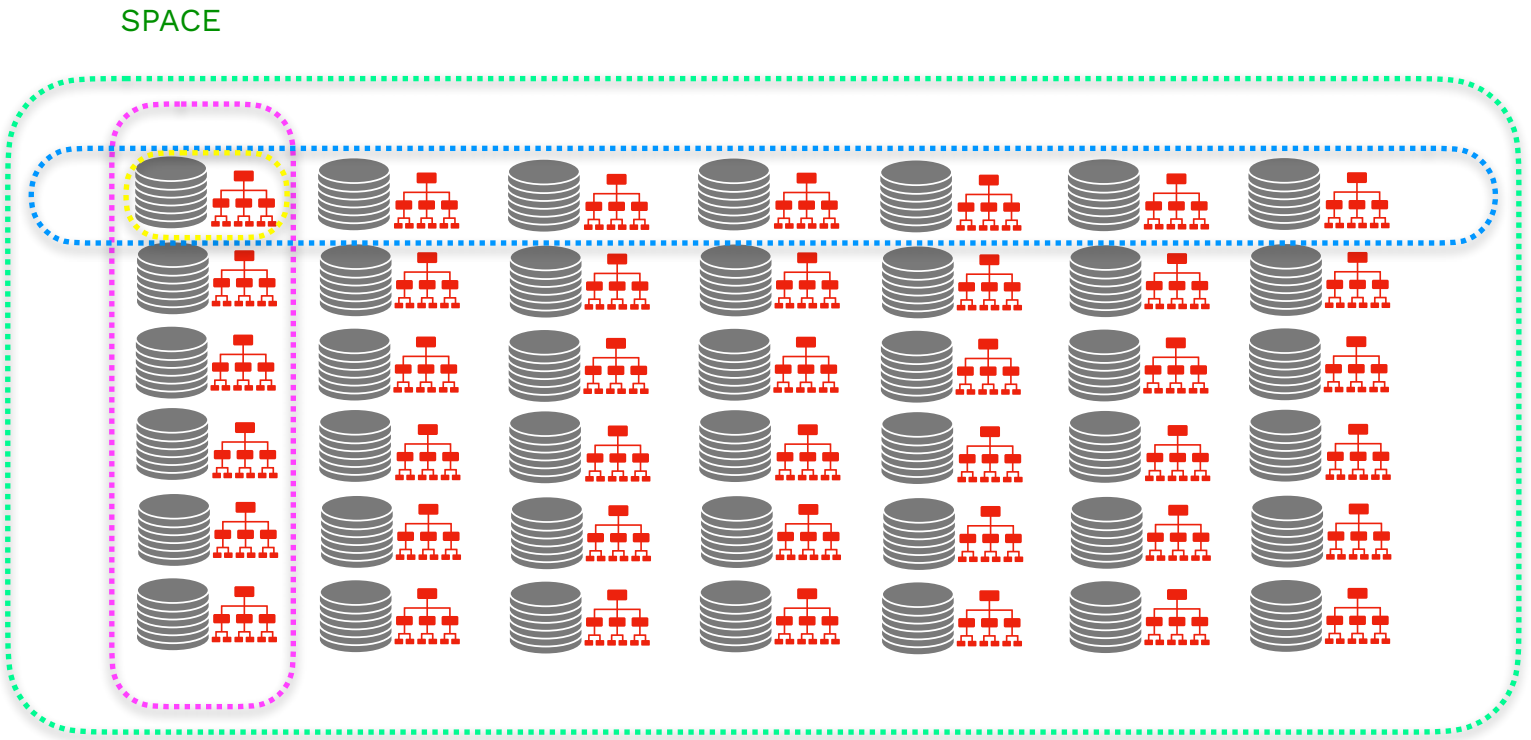
Client Protocols

CLI: root://
FUSEX: zmq://
+root://
http(s)://
(XrdHttp)

XRootD Framework



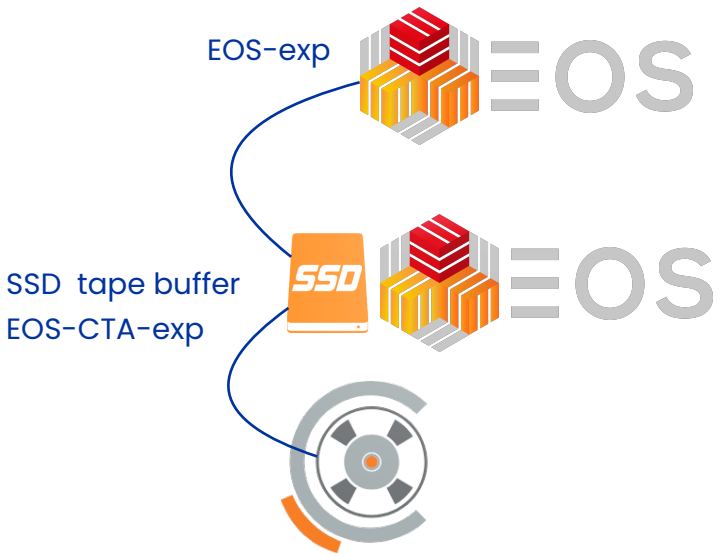
Architecture - Concepts



node: physical machine hosting filesystems
space: aggregation of groups = aggregation of filesystems
group: vertical aggregation of filesystems used for scheduling
filesystem: individual mounted device

Tape Write capabilities

- EOS now provides as well tape archive functionality
- EOS is natively used as a namespace and disk pool for the CERN Tape Archive (CTA)
- A pure SSD EOS instance with tape backend
- Conceived as a fast buffer to the tape system
- File residency on disk is transitional
- A tape copy is an offline file for EOS
- Intended to meet the requirements of Run3 and Hi-Lumi LHC



CERN
Tape Archive

```
  File: '/eos/archive/foo/bar'  Flags: 0000
  Size: 13106676363
Modify: Fri Jan 24 01:20:07 2020 Timestamp: 1579825207.000000000
Change: Fri Jan 24 01:20:07 2020 Timestamp: 1579825207.000000000
 Birth: Thu Jan  1 01:00:00 1970 Timestamp: 0.000000000
  CUid: 120682 CGid: 2766 Fxid: 6ae9383e Fid: 1793669182 Pid: 1776959392 Pxid: 69ea3fa0
XStype: adler    XS: 71 d5 9f 21    ETAGs: "481484404783316992:71d59f21"
Layout: replica Stripes: 1 Blocksize: 4k LayoutId: 00100012 Redundancy: d0::t1
  #Rep: 1
TapeID: 1793669182 StorageClass: cast1
```

no.	fs-id	host	schedgroup	path	boot	configstatus	drain	active	geotag
0	65535	localhost	tape.0	/does_not_exist		off	nodrain	offline	

```
*******
```

EOS at CERN

Some operational numbers

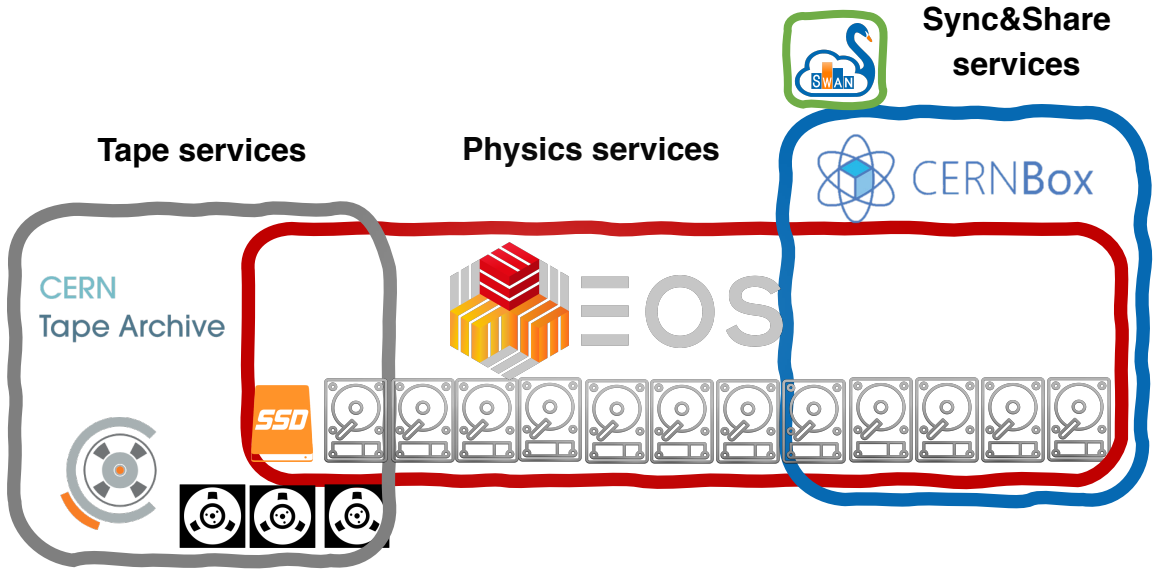
EOS at CERN

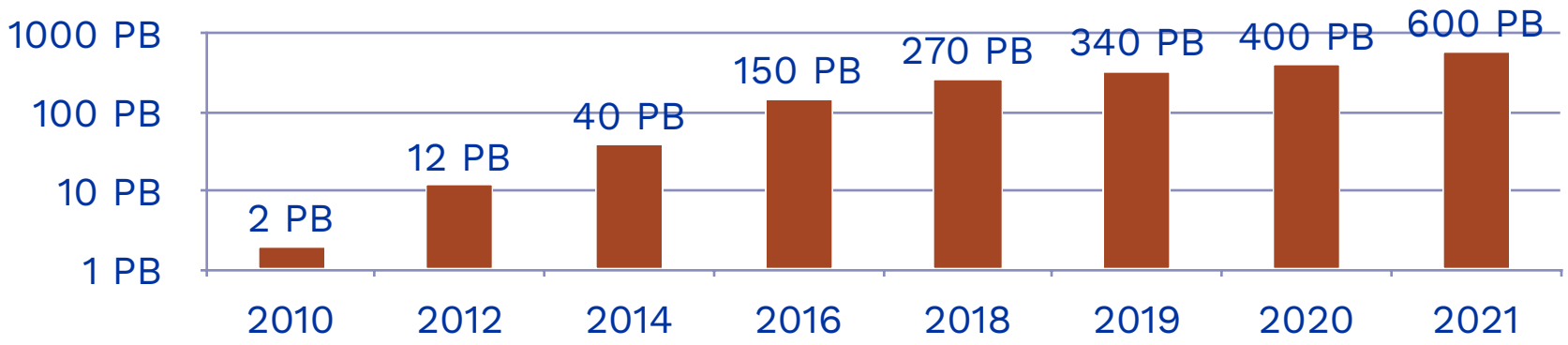
Total Managed Space
600 PB

Files Stored
~7 Billion

 # Storage Nodes
~1600

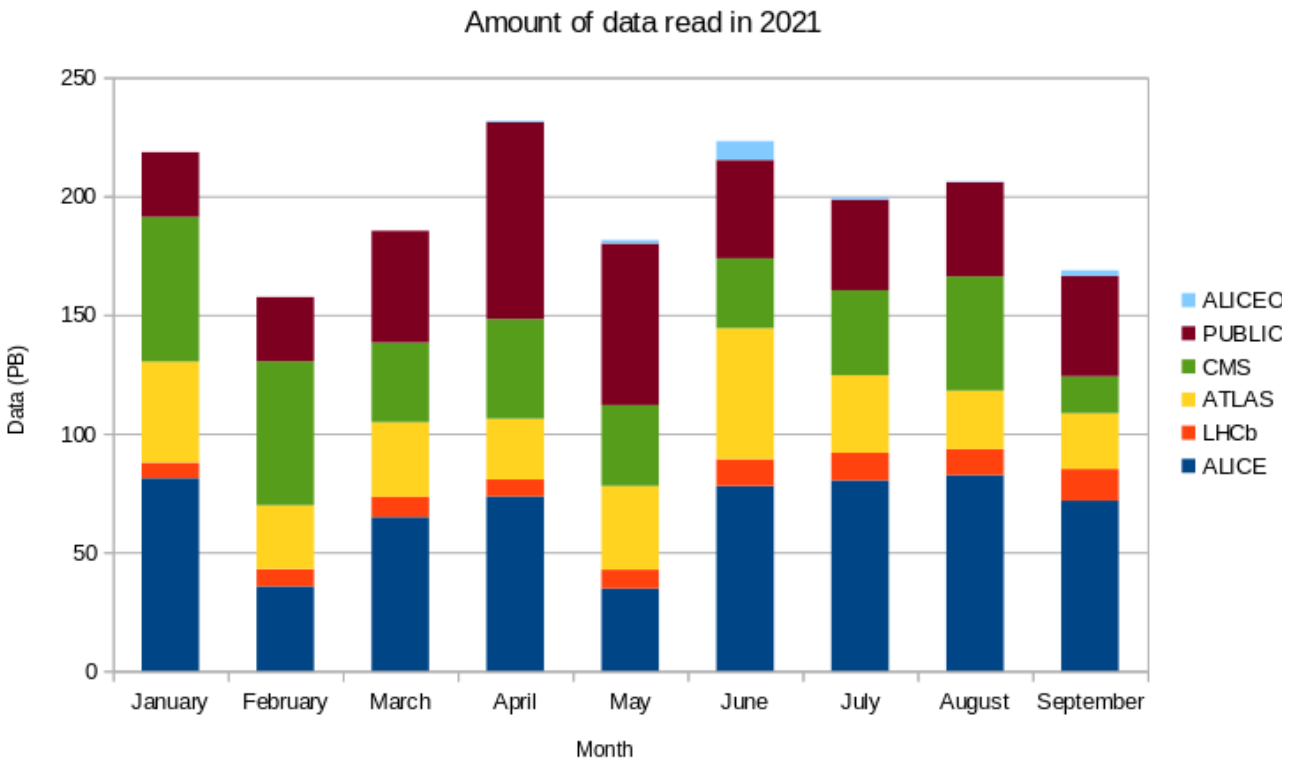
 # Disks
~80000



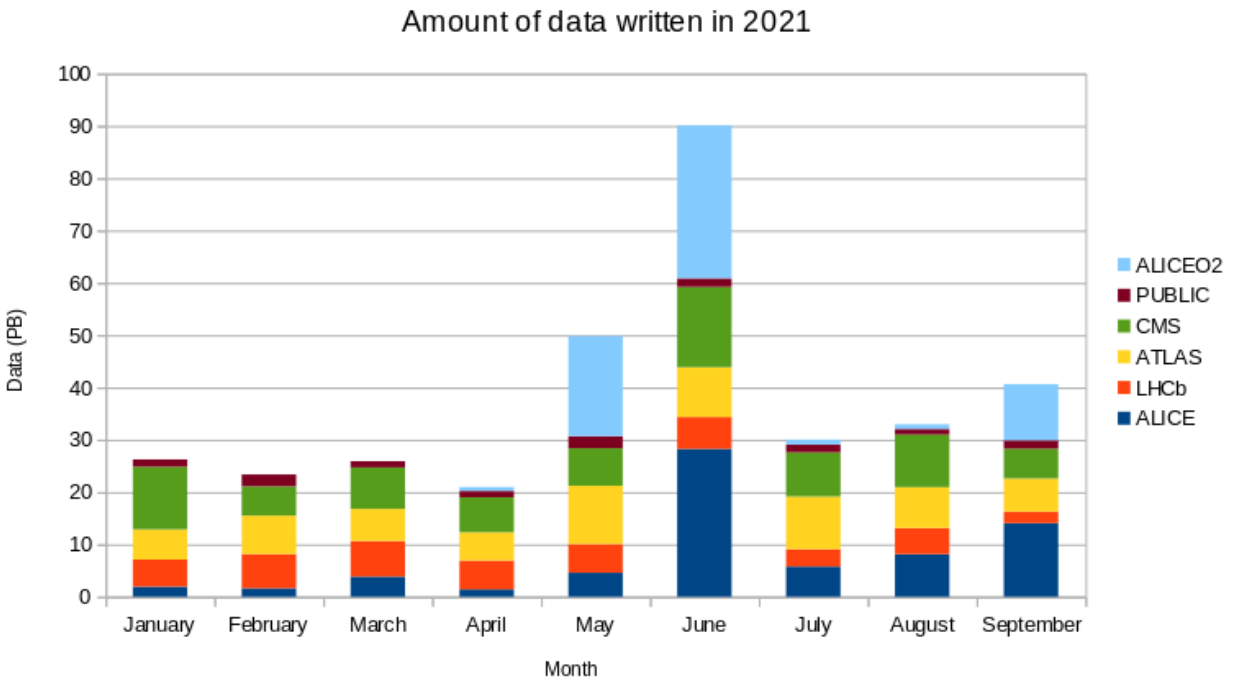


2021 in Data

~1.8 EB Read

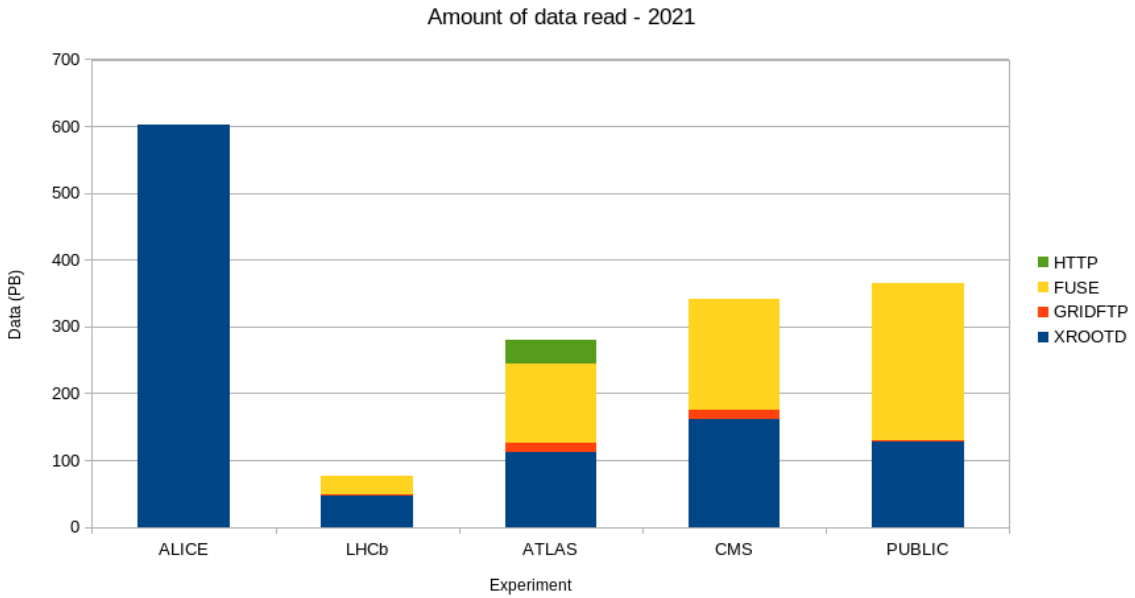


~350 PB Write

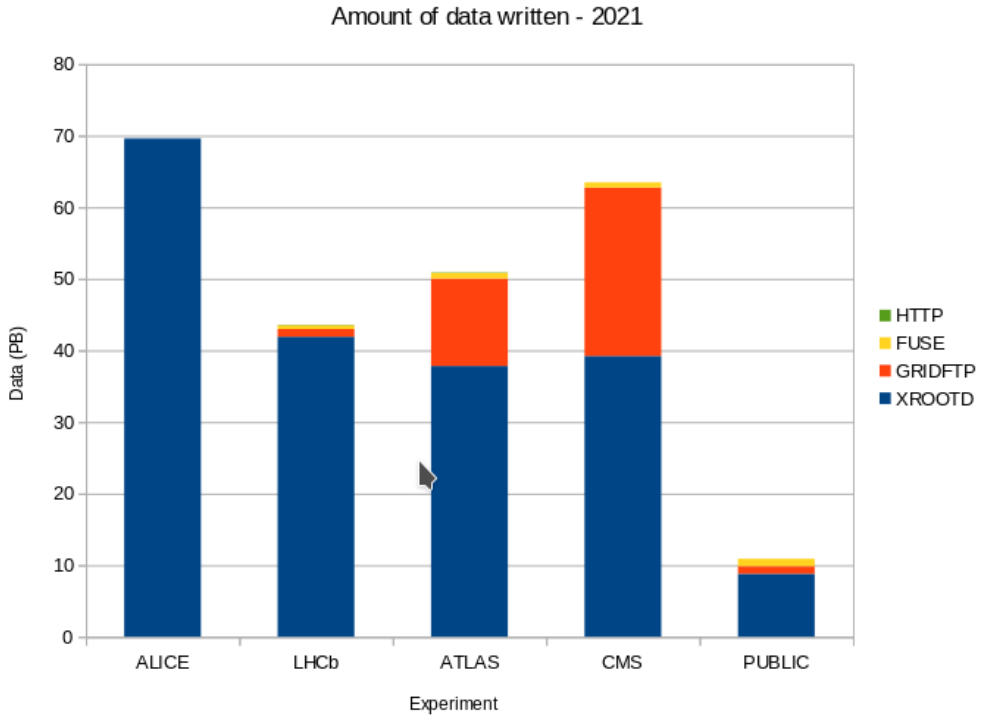


Data I/O by protocol

	XRootD (read)	GridFTP (read)	EOS-FUSE (read)	HTTP (read)
[PB]	1072	28	550	36

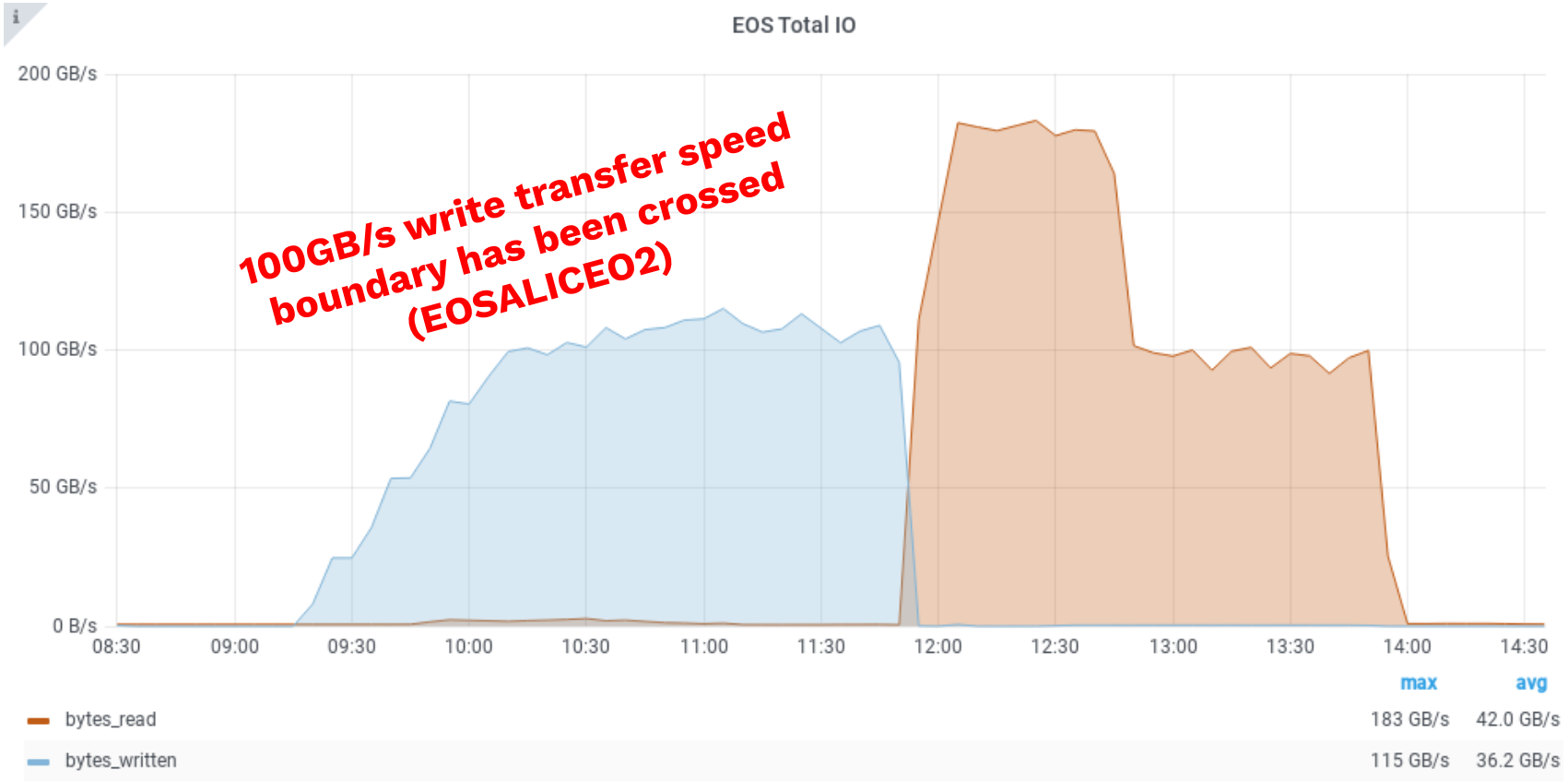


	XRootD (write)	GridFTP (write)	EOS-FUSE (write)	HTTP (write)
[PB]	198	38	3	0



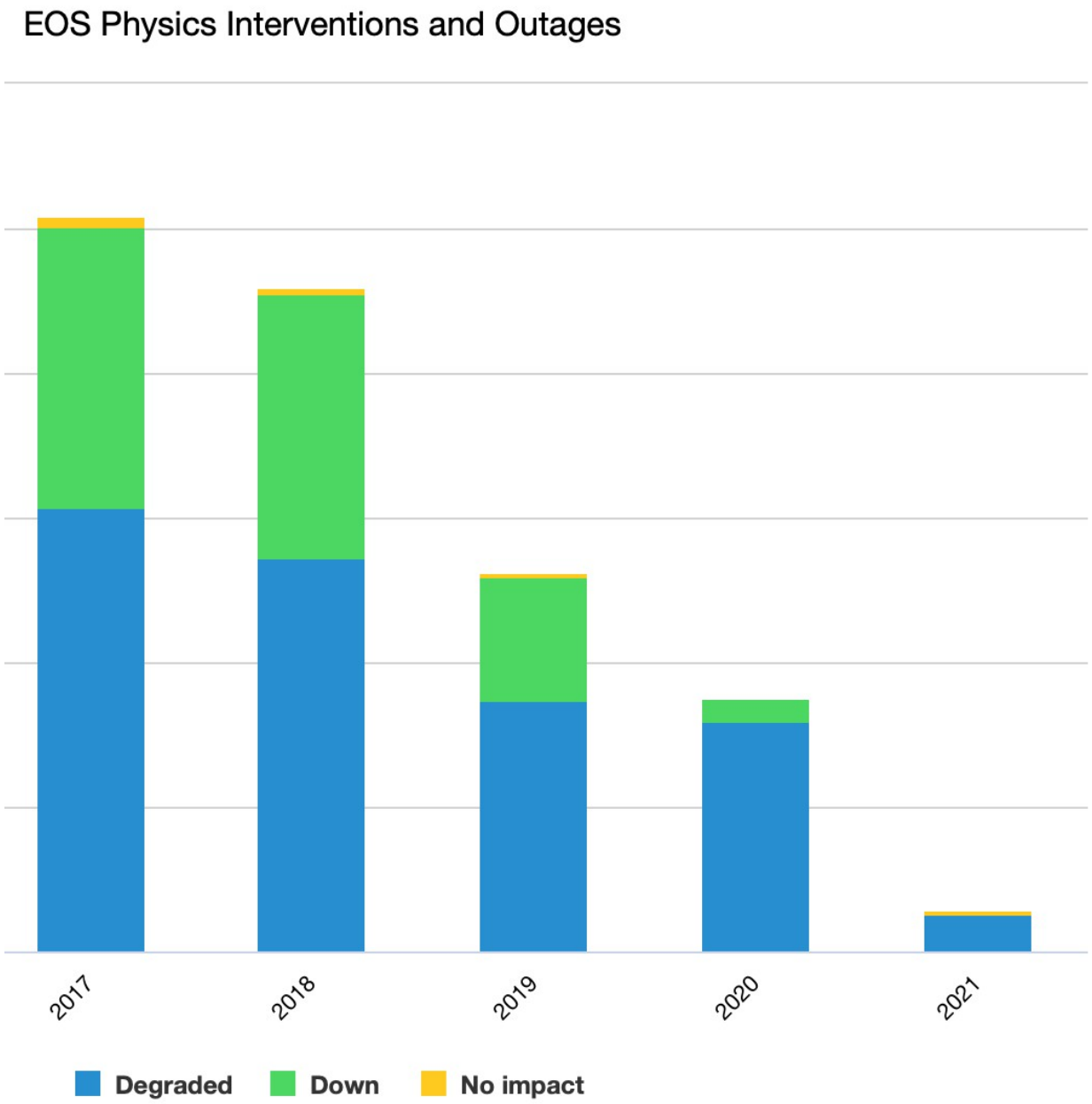
* HTTP-TPC traffic not tagged properly (grain of salt needed)

Achievements: BW



Availability

- Downtimes have almost disappeared
- Degraded issues are usually network problems and sometimes user overloads

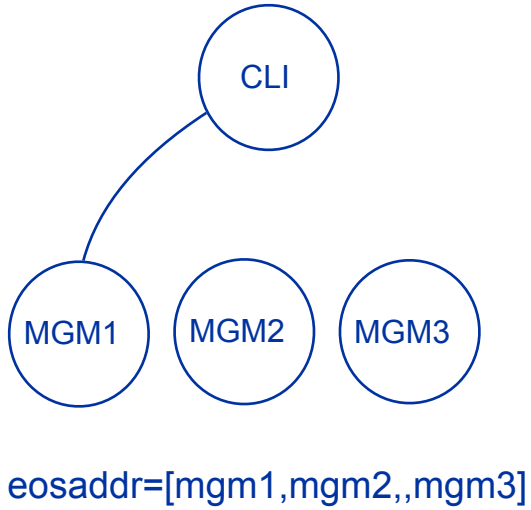


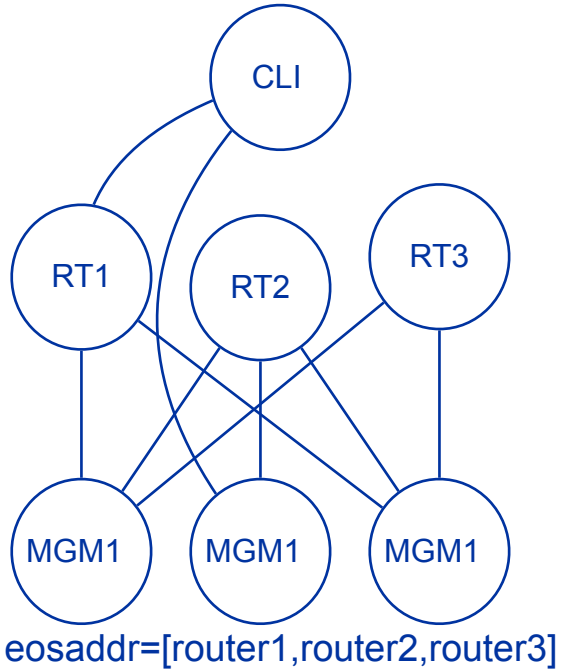
EOS 5

New features

EOS 5

- Uses **XROOTD5** as the framework; brings many new features:
  - Encryption (both data & metadata can be negotiated by client/server)
  - Extended interfaces for security, object etc
- Client features
  - *Redirect collapse:* Easier HA without need for DNS LB
  - *XrdEc:* Client side Erasure Coding
  - Declarative API, ease of async. programming
  - Kernel buffer support, vector writes
  - Memory consumption improvements in TPC transfers





EOS5: Future features & deprecations

Deprecations

- MQ: merging the functionality into MGM, QDB
- EOSD: deprecating in favor of EOSXD
- In memory namespace deprecation
- LevelDB: reusing QuarkDB wherever possible
- Move config to QDB, drop conf file

Features

- Simplified Scheduler
- Namespace API/locks refactoring
- EOSXD improvements

Packaging

QuarkDB will be a part of EOS release process. The package will be called eos-quarkdb

Roadmap

Roadmap

XRootD5
deprecations
new configuration

EGI release
X5 **HA**
eosxd fixes
share ACLs
CTA HTTP **API**

drop MQ, LevelDB, eosd
drop libmicrohttpd
simple **scheduler**
namespace api/locking **refactoring**

EOS5: ToDo List

- **Major**
  - all **balancers** refactored **with TPC**
  - XRootD native erasure coding **XrdEc** [comparison]
  - EC updates using **clone range**
- **Minor**
  - support for **multiple checksums** per file [Q4]
  - **OAUTH2** streamlining wrt **WLCG** specification [Q4]
  - **dynamic EC** - drop or adopt [decision Q4]
  - **documentation** & **web** update for **EOS**5 [Q4]

Deployment plans

- Straightforward upgrade: No breaking EOS4 -> 5 upgrade procedures; rolling update as usual
- Already rolled out in limited test EOS instances - PPS, Pilot, AMS
- Planned for all physics instances before Run 3
- Transparent upgrade for clients - expected on FUSEX clients later this year.

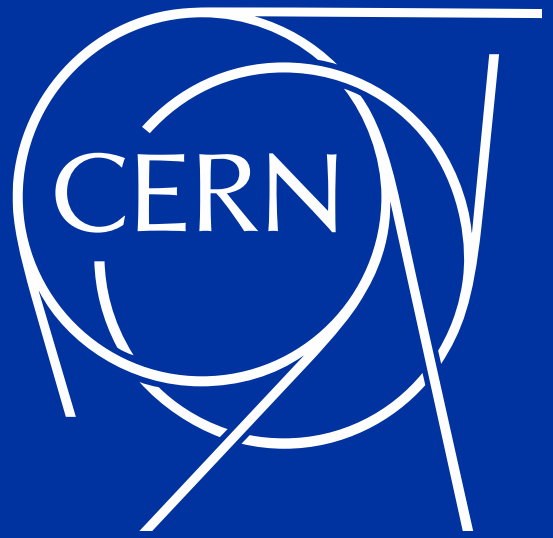
Features: Token Support

- WLCG tokens: supported for XROOTD/HTTP-TPC
- OIDC/OAuth2: supported for /eos/, uses server side mappings similar to grid-map files
- EOS tokens:
  - supported for all protocols ie. eos://, grpc://, https://, cli
  - optionally may contain ACL entry for directory/subtree and an optional identity

Support

- EOS has solid support from CERN
  - crucial software for physics data storage
  - crucial for user data storage CERNBOX
  - crucial for archival storage CTA
- EOS services at CERN run with best effort support only
- EOS has an increasing community and is deployed in tens of storage installations world-wide
  - RAL & CNAF exploring EOS and CTA
  - many WLCG Tier-2 sites
  - other institutes JRC, AARNet
- Business support provided by COMTRADE 360
  - Windows Native client
  - Ad-hoc features

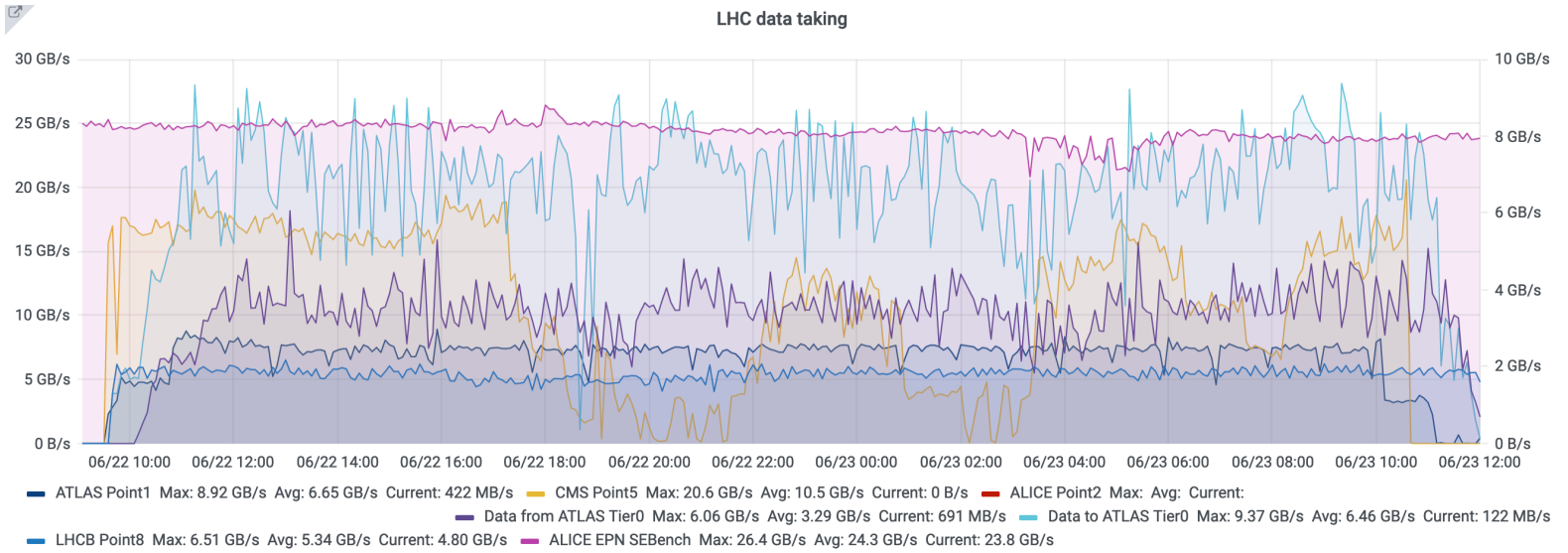
GIT: https://gitlab.cern.ch/dss/eos
Web: https://eos.web.cern.ch
Community forum:
Web: https://eos-community.web.cern.ch
Email: eos-community@cern.ch
Documentation: https://eos-docs.web.cern.ch/eos-docs
Support: eos-support@cern.ch


CERN

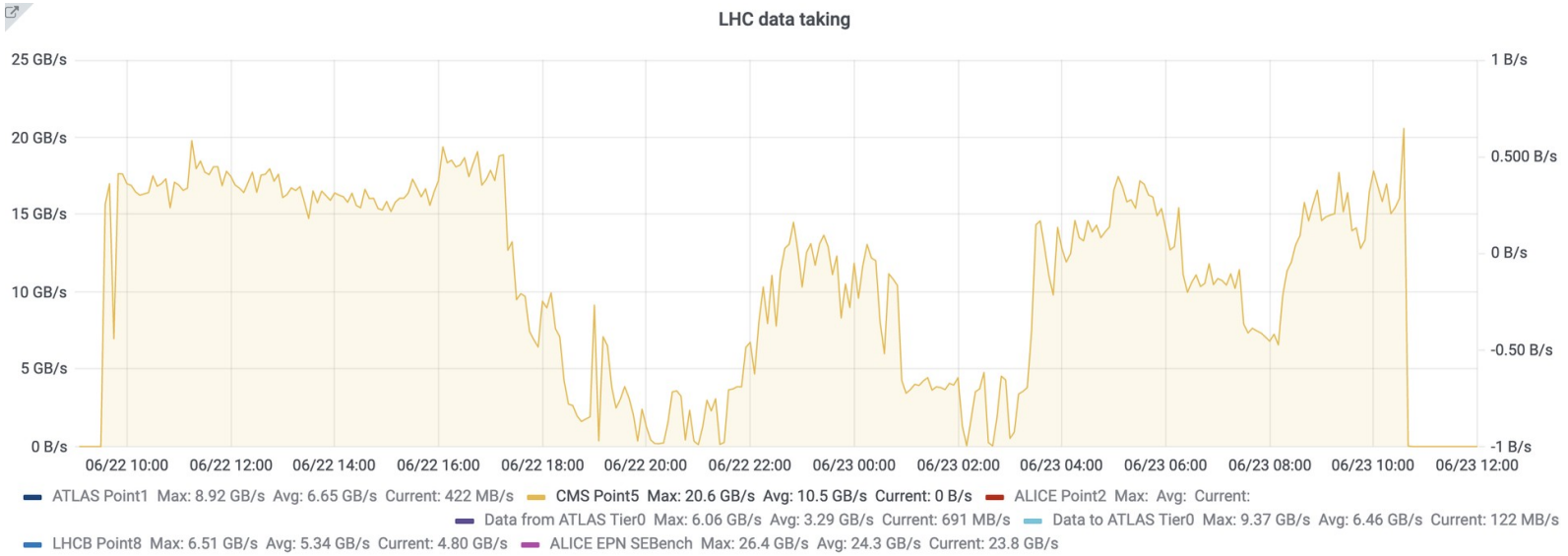
Extra Content

Achievements: Data Challenge

Bandwidth requirements for experiments were reached without interfering with others



CMS overcommitted EOS resources - resulted in hotspots (QoS feature*)



Feature: QoS

- QoS in EOS is configurable by space or even by directory
  - space is a grouping of hardware e.g. SSD space or EC space
  - a forced space is configured on a directory basis
- QoS in EOS* is configurable by application
  - IO streams are tagged automatically (eoscp, gridftp, fuse) or manually using CGI '*eos.app=myapp*'
  - IO streams
    - can be bandwidth limited (max MB/s)
    - can define disk IO scheduler policy (low...high, realtime)

Feature: QoS transitions

- QoS transitions can be triggered manually or automatically
  - manually by converting a file layout e.g. 2 replica to erasure coded 10+2
  - automatically by using a space policy or workflow events
    - e.g. CTA stages files back for ALICE on an SSD pool and on close a QOS transition is triggered to convert single replica files on SSD to erasure coded files on a larger HDD space
- files can be converted to a new layout by age, size, suffix policies, examples:
  - e.g. all files younger than 3 month stay on SSD, older files move to HDD
  - Files greater than 1 GiB and older than 3 days are converted from 2 replica to EC 10+2